The Use of Receiver Operating Characteristic (ROC) Curves as a Tool to Assess Noise in the Targeted Sequencing of Forensic Short Tandem Repeat (STR) Markers

Sarah Riman[1], PhD; Hari Iyer[2], PhD; Lisa Borsuk[1], MS;
Peter M. Vallone[1], PhD

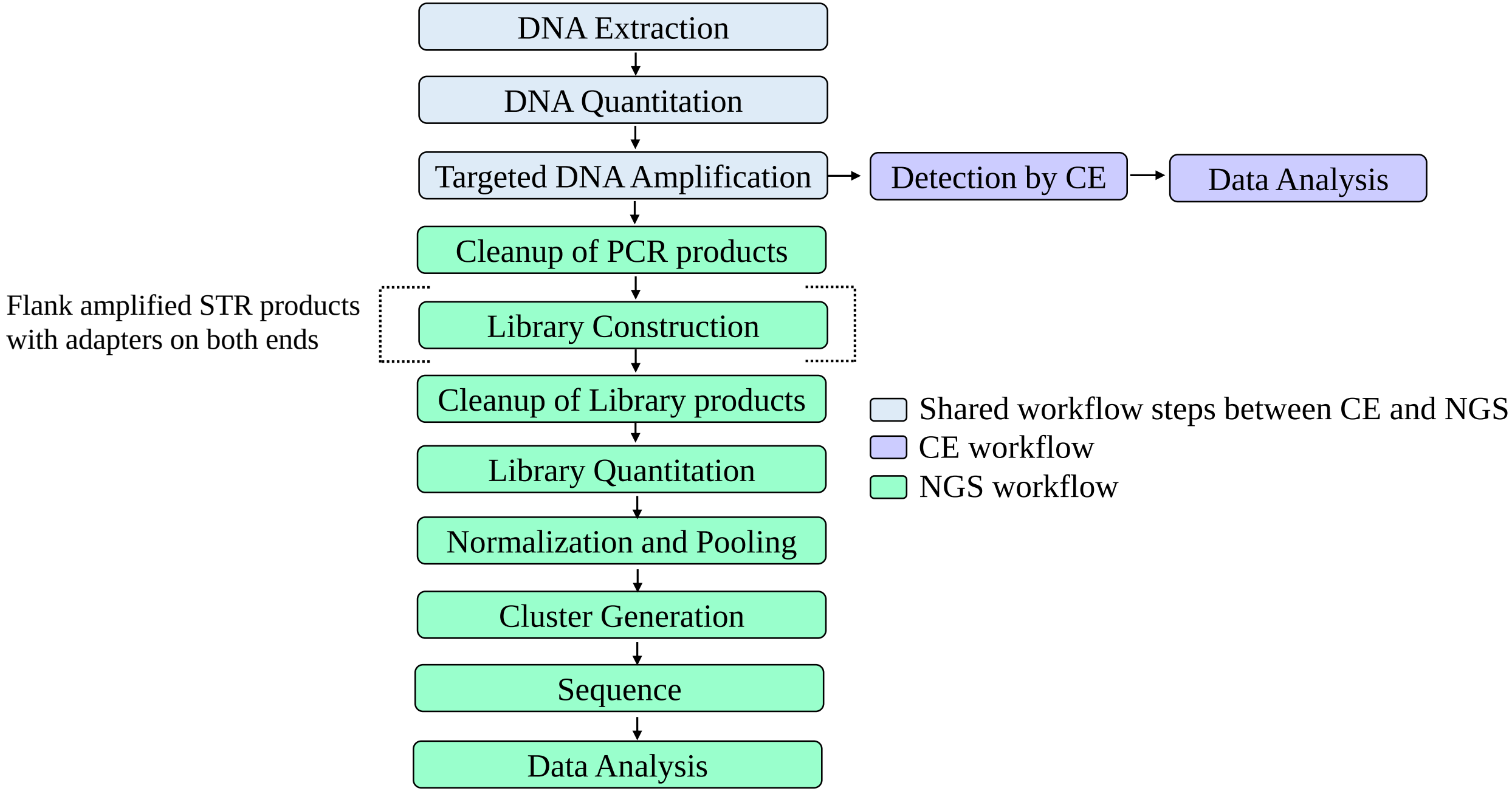
[1]Applied Genetics Group, National Institute of Standards and Technology
[2]Statistical Design, Analysis, and Modeling Group

Overview

- Comparison of conventional CE versus NGS-STR genotyping workflows
- Comparison of conventional CE versus NGS-STR data analysis
- Scope of the work
- Discussion of a sensitivity study consisting of single-source DNA profiles generated by targeted sequencing

Comparison of conventional CE versus NGS-STR genotyping workflows

Comparison of conventional CE versus NGS-STR genotyping workflows

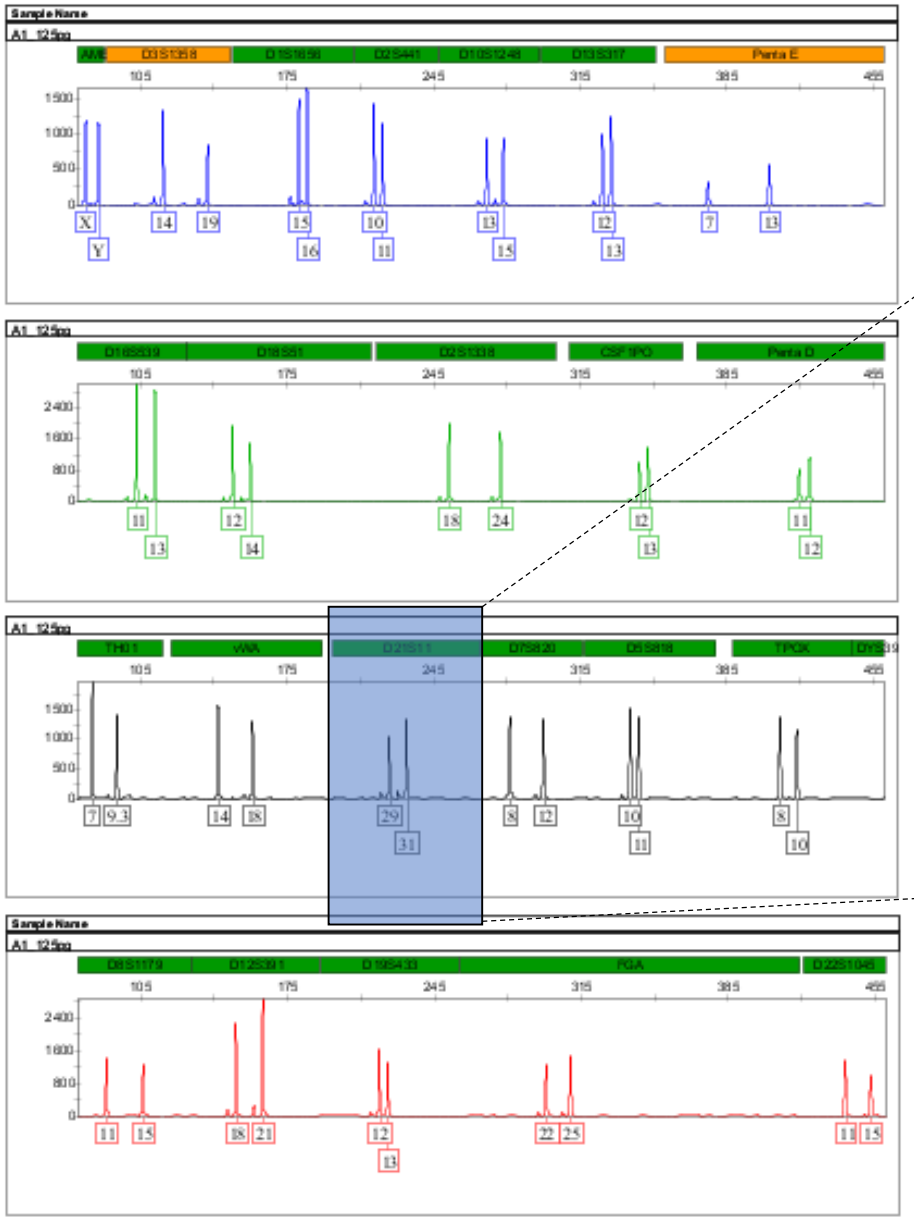


- Targeted sequencing of STR markers relies on the PCR-amplification process

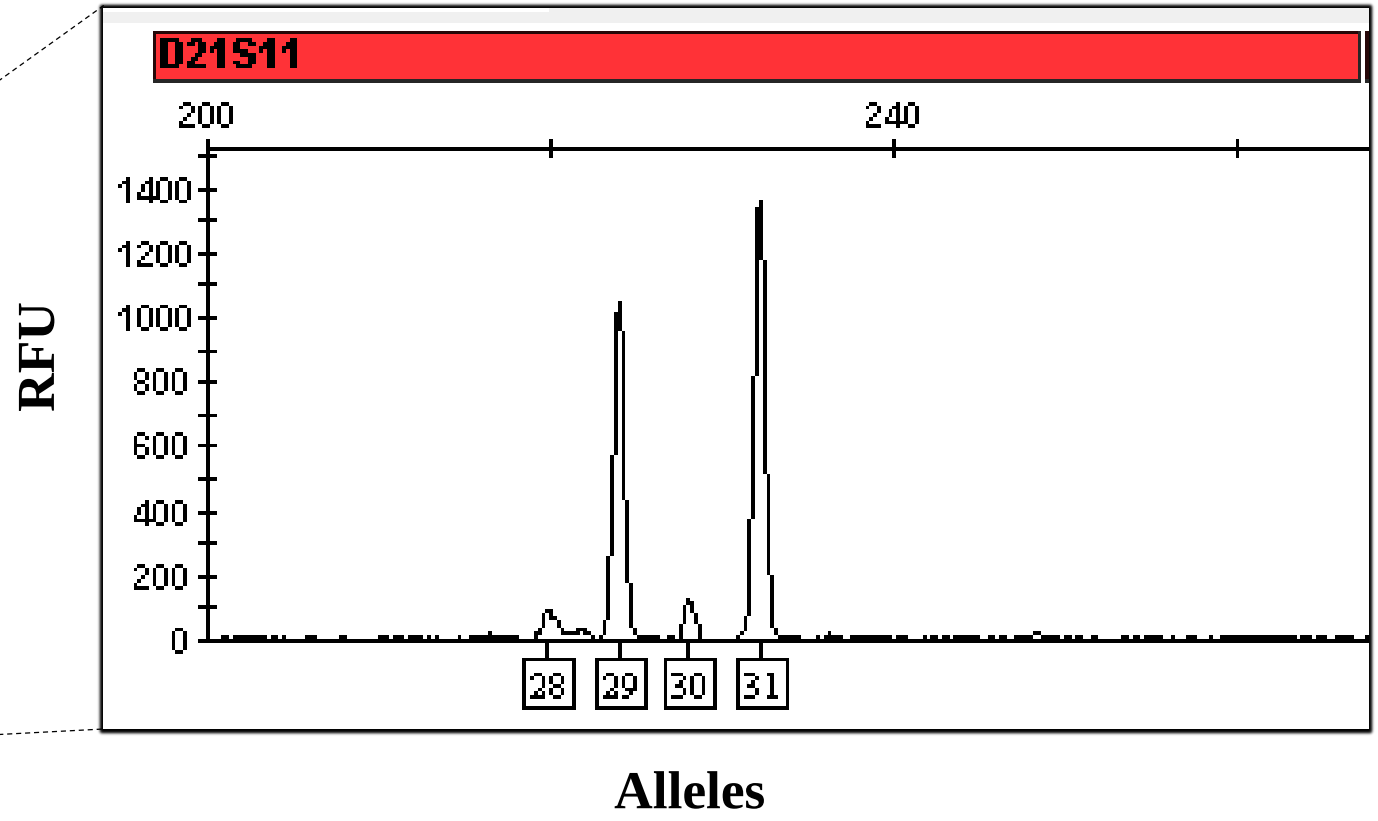
- NGS-STR profiles are susceptible to: **stochastic variation, signal noise, stutter artifacts, heterozygote imbalance,** and **allelic drop-out/in**

Comparison of conventional CE versus NGS-STR data analysis

Data Analysis by CE

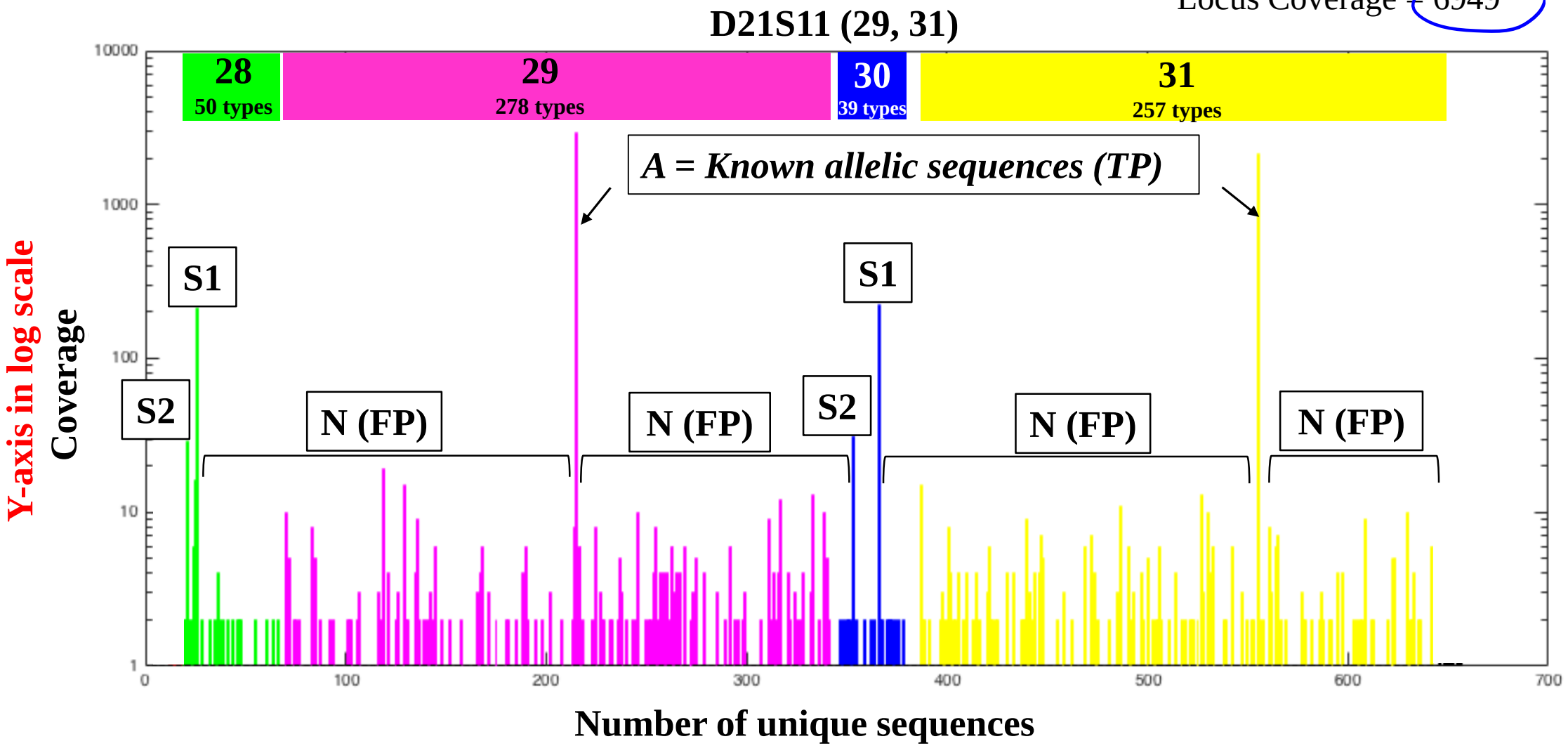


CE-D21S11



Data Analysis by NGS

Total number of unique sequences = 646
Locus Coverage = 6949



A = Known allele sequences (true positives: TP)

S1 = Back stutter of the LUS of the basic repeat motifs within an allelic sequence

S2 = Back stutter sequences not attributed to S1

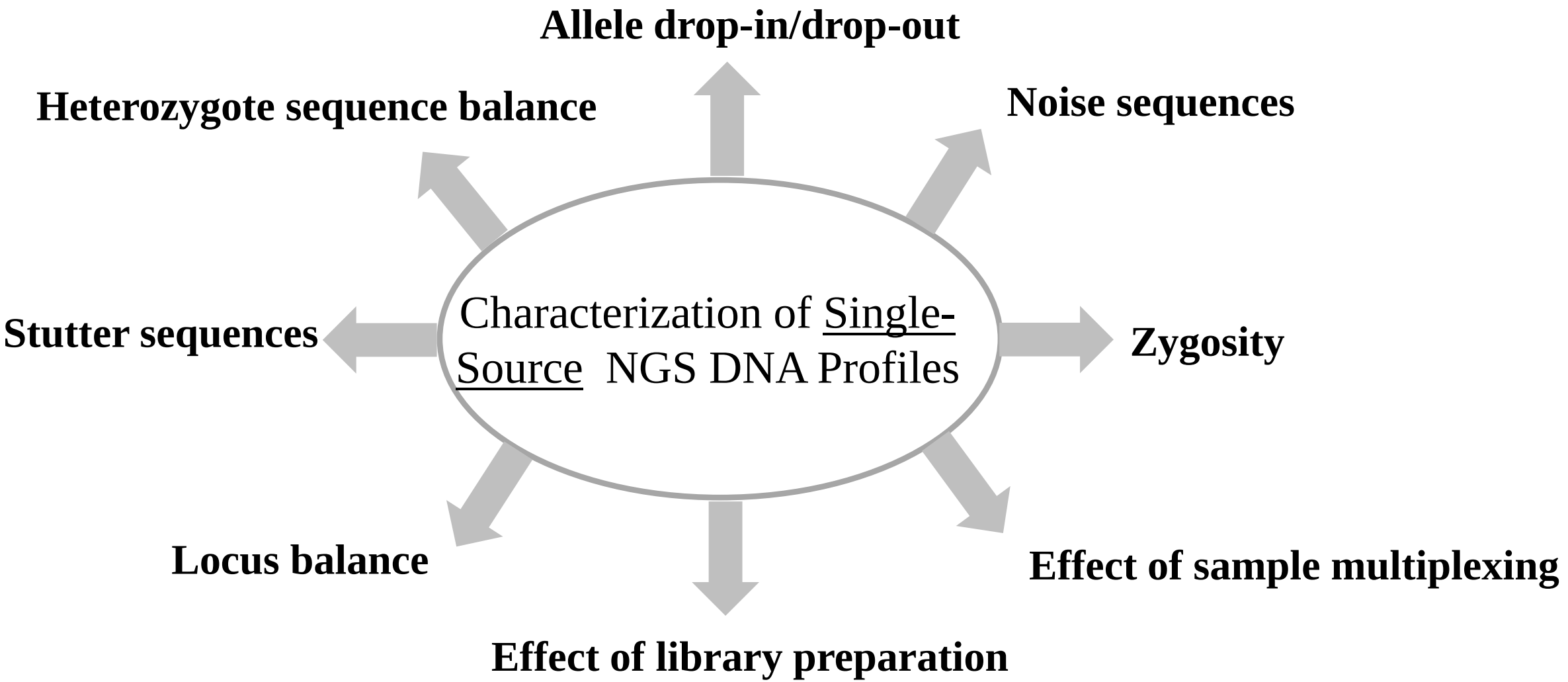
N = Noise sequences (false positives: FP)

What needs to be evaluated to:

Apply NGS-STR genotyping to forensic casework

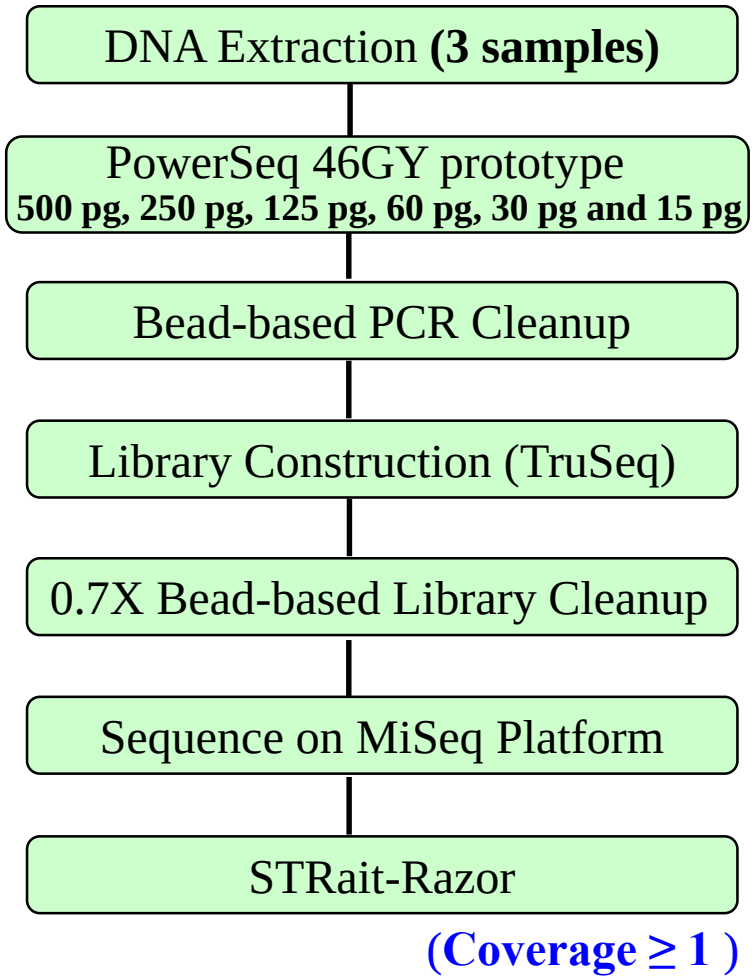
Establish probabilistic models for interpreting NGS-STR profiles?

We need to generate and characterize NGS-STR data



Sensitivity study of single-source DNA profiles generated by targeted sequencing

NGS sensitivity experimental design



PowerSeq 46GY System Prototype									
	1	2	3	4	5	6	7	8	9
A	500 pg	500 pg	500 pg	500 pg	500 pg	500 pg	500 pg	500 pg	500 pg
B	250 pg	250 pg	250 pg	250 pg	250 pg	250 pg	250 pg	250 pg	250 pg
C	125 pg	125 pg	125 pg	125 pg	125 pg	125 pg	125 pg	125 pg	125 pg
D	60 pg	60 pg	60 pg	60 pg	60 pg	60 pg	60 pg	60 pg	60 pg
E	30 pg	30 pg	30 pg	30 pg	30 pg	30 pg	30 pg	30 pg	30 pg
F	15 pg	15 pg	15 pg	15 pg	15 pg	15 pg	15 pg	15 pg	15 pg
G	**Sample A**			**Sample B**			**Sample C**		
NGS Workflow									

How should noise thresholds be set for NGS?

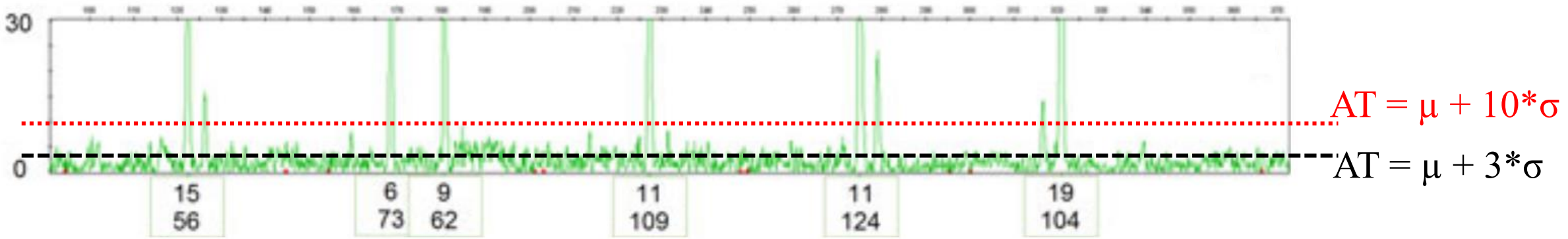
CE analytical threshold is most commonly determined by:

$$AT_{M1} = \bar{Y}_{bl} + ks_{bl}$$

- *k = Numerical factor (e.g. k=3)*

Average RFU signal

STDEV of the signal



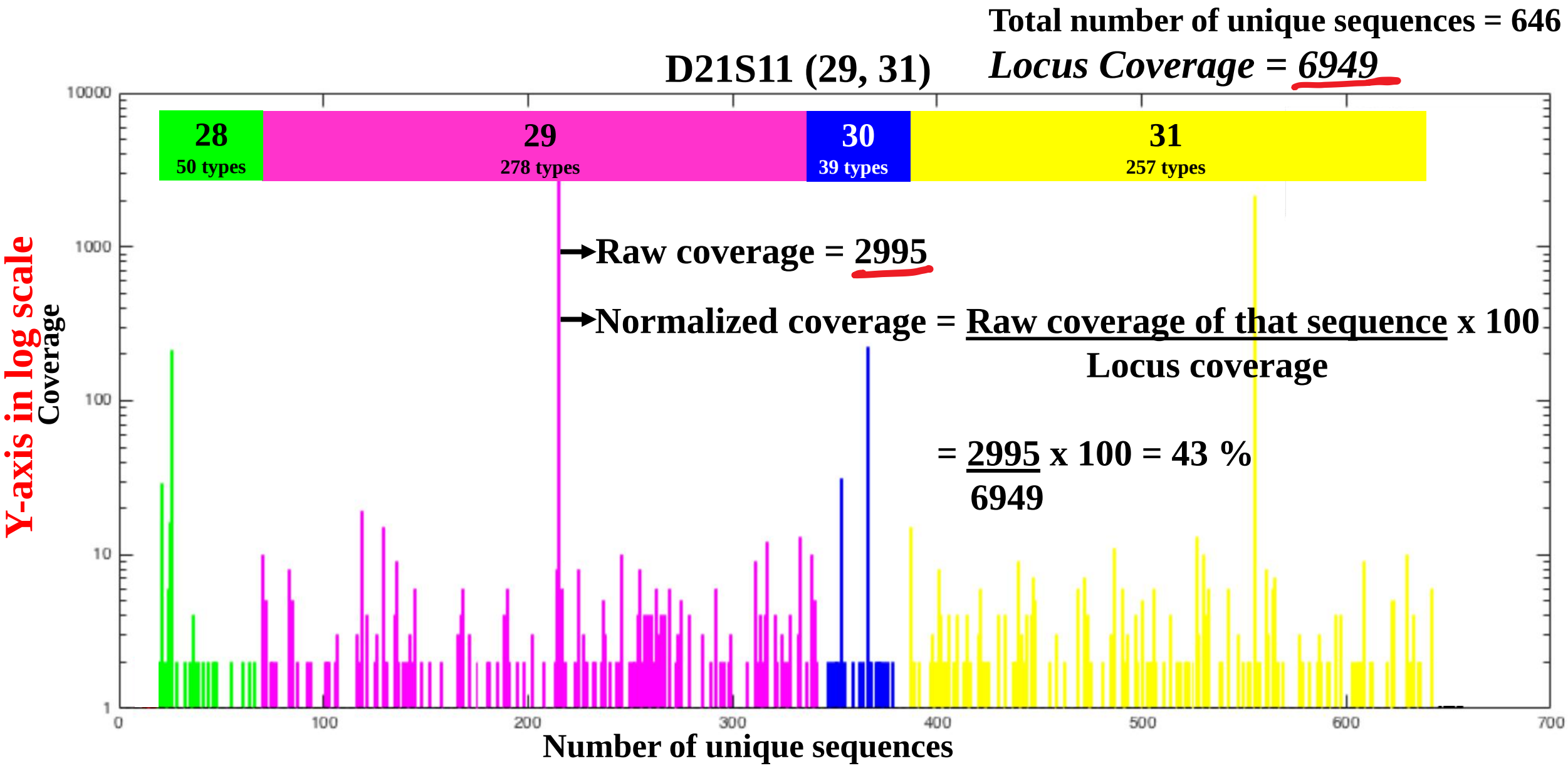
- **Dye-dependent**
- **Peak height (RFU)**

J. Bregu, D. Conklin, E. Coronado, M. Terrill, R.W. Cotton, C.M. Grgicak, Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis, Journal of forensic sciences 58(1) (2013) 120-9.

How to set noise thresholds for NGS?

CE-noise thresholds	NGS-noise thresholds
Peak Height (RFU)	Raw coverage or normalized coverage ?
$\mu + 10*\sigma$	?
Dye-dependent	▪ Protocol-dependent? ▪ DNA amount-dependent ? ▪ Locus-dependent ? ▪ DNA amount and locus-dependent ?

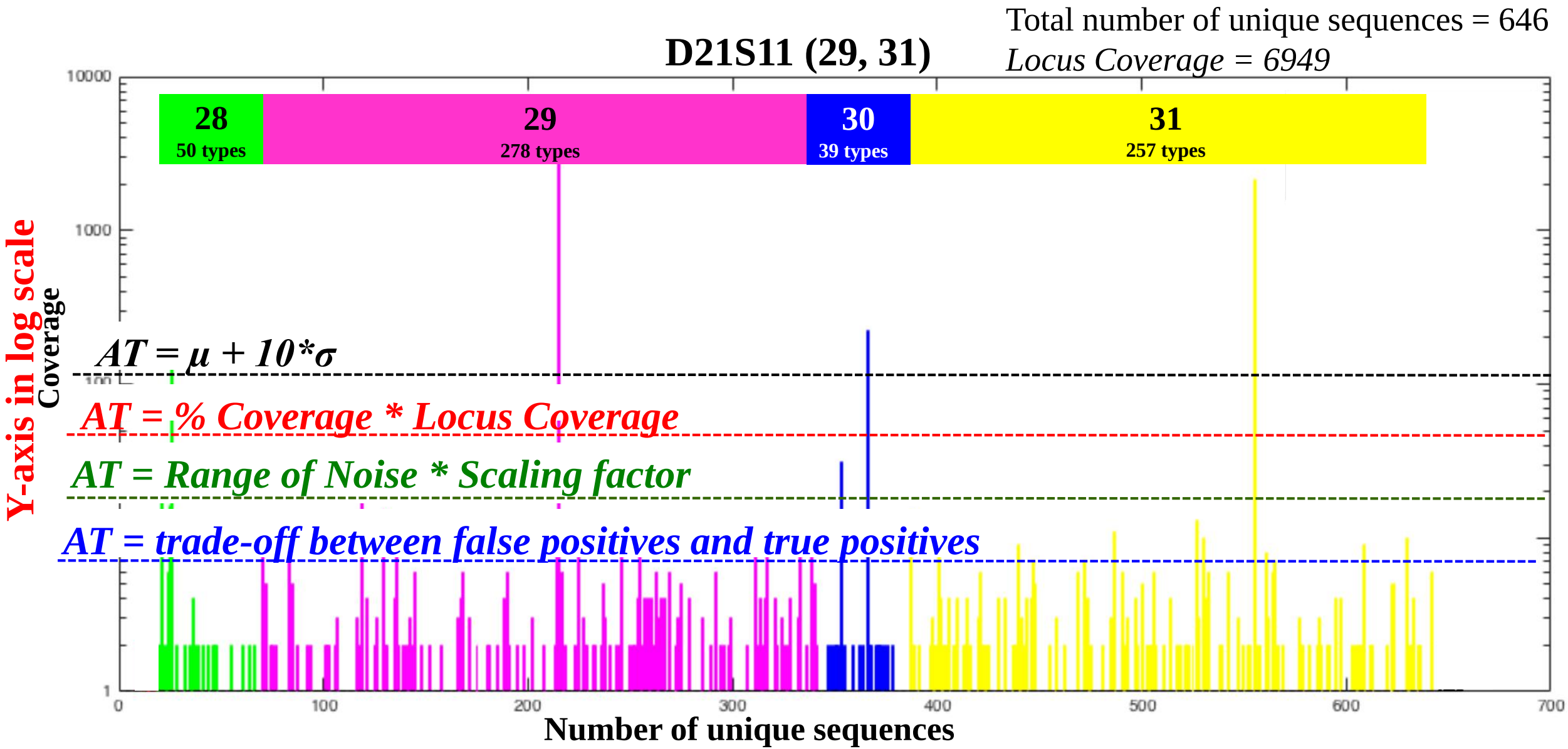
Raw coverage and normalized coverage



How to set noise thresholds for NGS?

CE-noise thresholds	NGS-noise thresholds
Peak Height (RFU)	Raw coverage or normalized coverage ?
$\mu + 10*\sigma$	?
Dye-dependent	▪ Protocol-dependent? ▪ DNA amount-dependent ? ▪ Locus-dependent ? ▪ DNA amount and locus-dependent ?

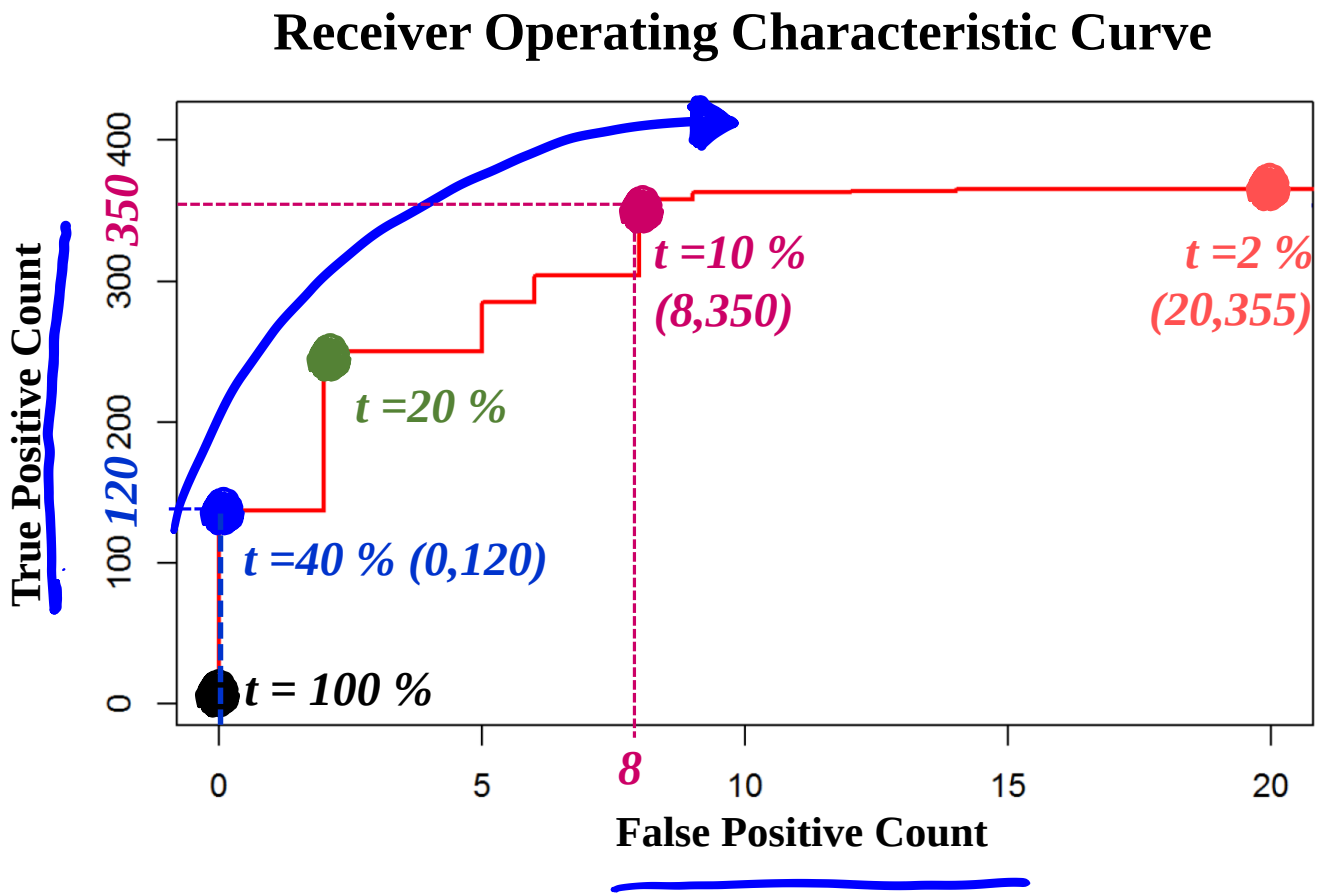
How to set noise thresholds for NGS?



Evaluating the tradeoff between the allelic (true positives) and noise sequences (false positives) using Receiver Operating Characteristic (ROC) Curves

Receiver Operating Characteristic (ROC) Curve

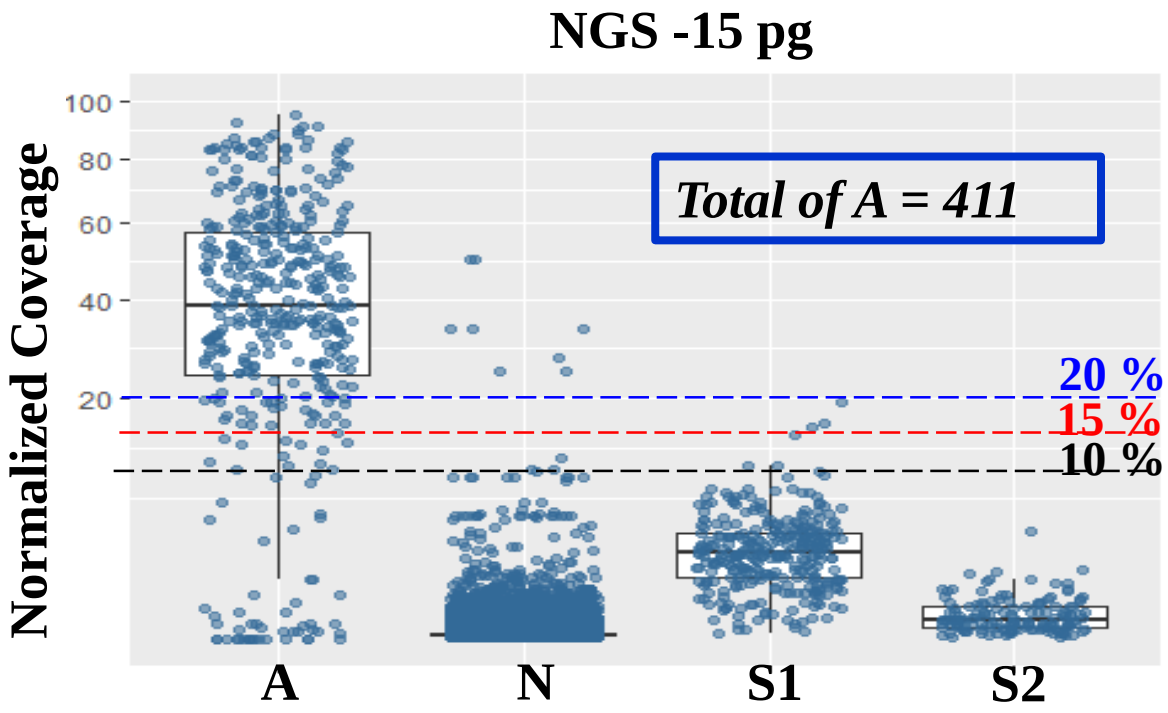
- **True positive count on y-axis**
  - Count of known allelic sequence
- **False positive count on x-axis**
  - Count of noise sequences
- **Captures all the noise thresholds (t) simultaneously**
- **High noise threshold**
  - Decrease in detecting known alleles
  - Decrease in drop-in
- **Low noise threshold**
  - Increase in detecting known alleles
  - Fewer drop-outs
  - Increase in drop-in

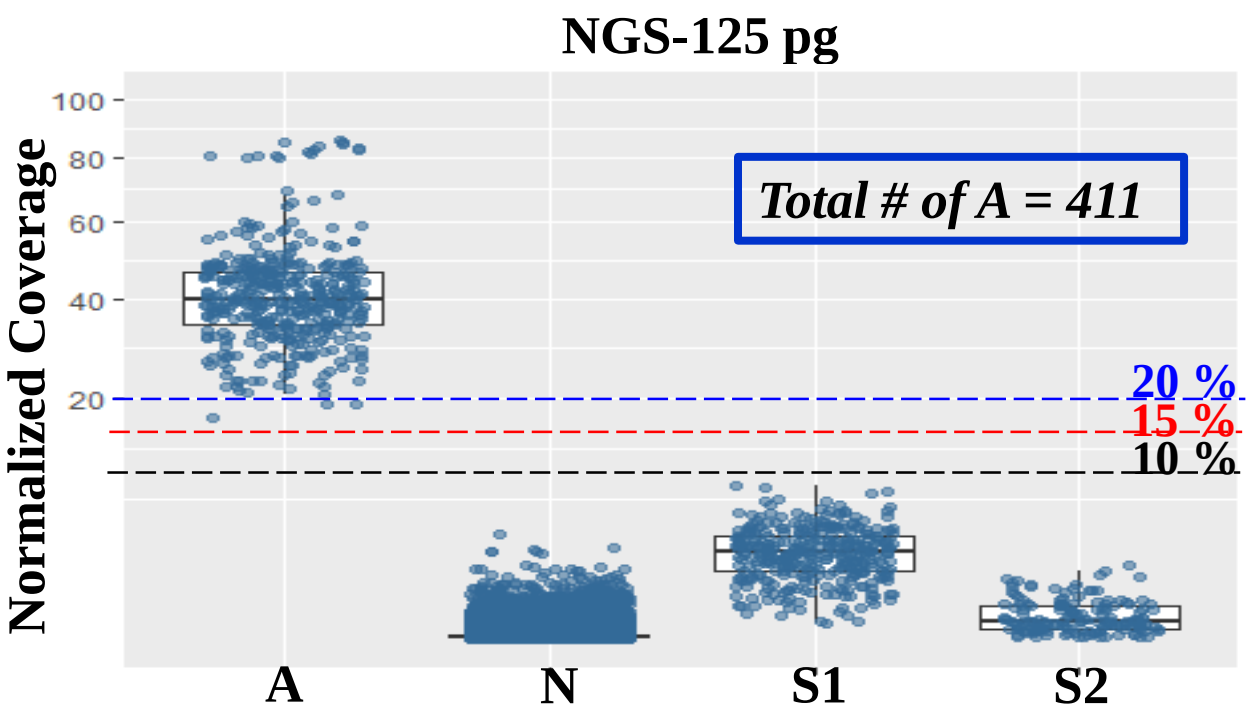


How to set noise thresholds for NGS?

CE-noise thresholds	NGS-noise thresholds
Peak Height (RFU)	Raw coverage or normalized coverage ?
$\mu + 10*\sigma$	?
Dye-dependent	▪ Protocol-dependent? ▪ DNA amount-dependent ? ▪ Locus-dependent ? ▪ DNA amount and locus-dependent ?

Impact of DNA Template Amount on the Distribution of Known Allele, Stutter, and Noise Sequences





t	A	DO	N	S1	S2
10 %	363	48	11	6	0
15 %	350	61	8	3	0
20 %	327	84	8	0	0

t	A	DO	N	S1	S2
10 %	411	0	0	0	0
15 %	411	0	0	0	0
20 %	408	3	0	0	0

- As expected, improved discrimination between known alleles (A) and the remainder of the sequences (N, S1, and S2) is observed as the amount of DNA template increases.
- Values of 10 %, 15 %, and 20 % are ONLY used for illustrative purposes and not as recommended thresholds. Each lab should perform sensitivity experiments and establish a threshold for interpretational purposes.

Future work

- Generate and characterize a large number of single-source samples with varying quantity and quality
- Characterize stutter sequences
- Characterize noise sequences
- Evaluate and compare using raw coverage and normalized coverage noise thresholds
- Evaluate DNA amount- and/or locus-dependent noise thresholds
- Generate and analyze mixture samples for performance check

Disclaimer

Points of view in this presentation are mine and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Commerce.

NIST Disclaimer Certain commercial products and instruments are identified in order to specify experimental procedures as completely as possible. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of these products are necessarily the best available for the purpose.

Acknowledgement

NIST

Pete Vallone

Hari Iyer (Statistical Design, Analysis, and Modeling Group)

Lisa Borsuk

Becky Steffen

Erica Romsos

Katherine Gettings

Kevin Kiesler

Margaret Kline

Megan Cleveland

Promega

Doug Storts

Spencer Hermanson

Funding

NIST Special Programs Office: *Forensic DNA*

FBI Biometrics Center of Excellence: *Forensic DNA Typing as a Biometric tool*.

All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Contact: sarah.riman@nist.gov